Global and Local Constraints for Neural Models of NLP: A Thesis Proposal

Dávid Javorský
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
javorsky@ufal.mff.cuni.cz

Abstract

Insufficient training data is a typical issue for many NLP tasks. In this thesis proposal, we design a new method which allows to transfer the knowledge between models trained on tasks of different nature. We experiment with our approach in a particular type of situations where the user wants to impose some constraints on the output. A very simple way would be to filter the training data to satisfy the constraints, exacerbating the small data issue. Instead, we define two tasks: a *constraining task*, which in some way exhibits properties relevant for the desired constraint, and a *primary task*, i.e. the task that should be solved under the given constraints. We train a model for the constraining task and repurpose attribution techniques, which originally aim to explain models decisions, to identify constraint features. Then, we train a model for the primary task and use the collected attribution scores as an additional information source to integrate the constraints.

1 Introduction

The improvements of neural models recently surpassed our expectations in various areas of natural language processing. Concerning neural machine translation (NMT), the quality of models is in some cases comparable with humans' decisions or even better (Popel et al., 2020; Barrault et al., 2019). However, there is still a notable gap in the output quality for tasks which are too specific under some constraints, and for which we usually do not have enough training data, e.g. the translation of low resource languages (Weller-di Marco and Fraser, 2022). The dependence on large amounts of training data brings a new problem, the inherent bias towards facts and formulations exemplified in the data (Garrido-Muñoz et al., 2021; Hovy and Prabhumoye, 2021).

In many situations, the user of a trained model can provide not just the input but also additional specific pieces of information which crucially influence which outputs are desired and which are not, in various aspects. For example, considering a multi-modal NMT setting, the performance of models increases when providing more than one channel (Sulubacak et al., 2020). This gives a rather vague form of constraints. Such additional information can be also given as a simple one-dimensional feature: In speech translation, knowing the gender of a speaker is a critical bit of information when translating from a language which does not express gender very often to a language which requires this information for every verb. It is known that current translation systems still struggle with correct gender choice (Kocmi et al., 2020; Stanovsky et al., 2019). Similarly, knowing the presentation criteria in subtitling helps the machine translation model conveniently fit the output to predefined length or decompose some articulations into shorter units, compared to what the generally available parallel texts exhibit (Karakanta et al., 2020; Zhang et al., 2020). In both cases, the model is provided with constraints from the user and we expect the model to adhere to them.

We can view that some of these constraints are *global* in the sense that the whole run of the model during a given session should reflect them (speakers' gender choice in translation, translation length), some of these constraints are *local* in the sense that the model's previous output affects their values.

Constraints typically limit the use of available training data. In such circumstances, the quality of a neural model can be improved by pretraining the model on a more general domain and gaining the task-specific knowledge from the target domain afterwards. This procedure is known as transfer learning and was first introduced by Bozinovski and Fulgosi (1976). Thanks to this, the dependence on large amounts of the task-constrained training data for constructing neural models is reduced.

Transfer learning is a powerful way to share learned knowledge across several domains which are different but related. However, there are situations when the number of samples from the target domain is still insufficient or a suitable related dataset does not exist because of its high specificity. In such situations, transfer learning, as it is known and used, cannot be employed.[1]

Contribution In this thesis proposal, we propose an innovative (to our best knowledge) method of transfer learning that allows combining datasets of a different nature. It consists of two steps.

In the first step, we repurpose one of the attribution techniques (De Cao et al., 2020) originally introduced to explain the predictions of a model trained for a specific task. Attribution methods typically compute sentence-level scores for each input unit (token, word, segment), identifying the ones that contribute most to the decision of the given model. By targeting a task which strongly expresses some desired constraints, we believe to extract attribution scores that correlate well with the importance of the constraint features.

In the second step, we analyze how to effectively apply these attribution scores to the models which are trained for unconstrained, yet well data-covered tasks. Note that this is a limitation of this method: We suppose that such models are capable of accepting and working with such scores as additional inputs. We provide the analysis of two tasks: Shortening in NMT and Gender Bias in NMT. We describe both tasks and provide a summary of available datasets for their evaluation.

Outline In the next section, we present an overview of the related work: Model analysis and its interpretability, transfer learning, and two selected problems in NMT that we focus on. Section 3 describes our proposed method. Section 4 discusses our approach of addressing two known problems from the perspective of our proposed transfer learning method. Section 5 describes selected experiments and preliminary results. In Section 6, we conclude the proposal and present plans for the future work.

[1]An alternative is few-shot learning, a method that allows models to solve tasks providing only few examples or 'shots' (Sulubacak et al., 2020), which was first shown in language modelling (Brown et al., 2020). Since this behaviour is considered as an emergent property in Large Language Models (Wei et al., 2022), we limit ourselves to utilizing the few-shot learning approach because of the requirements on spacial and computational resources, which the usage of such large models brings.

2 Related Work

2.1 Model Analysis and Interpretability

Given the popularity and the number of useful applications of neural models in natural language processing, there is a need to be able to interpret the behaviour of such models. Comprehensive reviews by Madsen et al. (2022); Belinkov et al. (2020) present a number of approaches, such as probing classifiers, studies on language modelling or inference tasks, layerwise analyses, neuron activations or attention mechanisms.

Our work focuses on quantifying the importance of input units (tokens, words, segments) in the decisions computed by the neural network. A simple approach along these lines analyzes attention patterns in neural models (Clark et al., 2019). Alternative methods assign significance scores to input tokens, such as in attention-based attribution (Vashishth et al., 2019), back-propagation (Sundararajan et al., 2017) or perturbation-based techniques (Schulz et al., 2020; Guan et al., 2019). A recent approach addresses several limitations of previous methods and introduces a new way of obtaining attribution scores where the training is performed in a fully differentiable way (De Cao et al., 2020). The authors show the analysis of a BERT model (Devlin et al., 2019) finetuned on question answering and sentiment classification tasks. We use this technique to obtain so-colled *attribution scores* which capture constraints in the task, and which are further helpful for training the general model. A more detailed description of this approach is in Section 3.

2.2 Transfer Learning

Transfer learning is a research technique that focuses on sharing knowledge across different task. During past years, diverse approaches have been proposed. We describe a selection of possible directions with the focus on natural language processing.

Meta learning One option is meta learning, also called 'learning to learn', which is an approach of observing learning from a wide range of tasks and learning new tasks much faster than otherwise possible. It is also capable of learning on very few examples of the training data, e.g. in the context of low-resource translation when the model learns to adapt to low-resource languages starting from multilingual high-resource language tasks (Gu et al.,

2018). A comprehensive review is provided by Hospedales et al. (2021).

Continual learning Another direction is continual learning which represents learning on the set of tasks where the model observes, once and one by one, examples concerning a sequence of tasks (Lopez-Paz and Ranzato, 2017). Only the data from the current task are available and the tasks are assumed to be clearly separated. The difficulty lies in not forgetting past tasks but accumulating the knowledge. A detailed summary is given by van de Ven and Tolias (2019).

Multi-task learning In multi-task learning, multiple tasks are all presented throughout the training, not one by one. The challenge is to balance them well for the best performance in the desired ones: E.g training a many-to-many model on a number of weakly related tasks such as machine translation, constituency parsing, image captioning, sequence autoencoding (Subramanian et al., 2018; Luong et al., 2015). A comprehensive review is presented by Zhang et al. (2023).

Ensemble learning Ensemble learning is another approach to transfer learning: An ensemble method is a machine learning technique that combines several base models in order to produce one predictive model with better prediction quality. The summary of existing ensemble methods is described by Sagi and Rokach (2018).

A different categorization of transfer learning is from the perspective of supervision: Unsupervised approaches, which aim to model the dataset itself without any provided annotation (Devlin et al., 2019; Radford et al., 2018; Howard and Ruder, 2018); And supervised approaches, which utilize large annotated datasets (McCann et al., 2017; Conneau et al., 2017; Yang et al., 2017). Independently of our summary, a comprehensive review of transfer learning methods is presented by Alyafeai et al. (2020).

In our approach, compared to these transfer learning techniques, we use a separate task, training dataset and model, where the analysis of such outputs is used as an additional input for the general model. Other techniques employ either one model that is trained for several different tasks (continual, multi-task learning), or several models that are trained for a similar task (ensemble learning).

2.3 Shortening Machine Translation

The objective of shortening machine translation is to generate a translation whose length is smaller than the source or reference length. Shortening machine translation can be viewed as a monolingual text compression with a follow-up machine translation. One direction is to filter training data and use only sample pairs where the source is shorter than target sentence (Macháček et al., 2021).

More complex approaches include neural end-to-end models developed for neural machine translation, notably the Transformer model (Vaswani et al., 2017), by explicitly introducing length constraints to control the behavior of the decoder (Lakew et al., 2019; Kikuchi et al., 2016), or incorporating length information to positional embeddings (Takase and Okazaki, 2019).

Alternatively, text compression can be addressed from the perspective of length disentanglement (Thompson and Post, 2020). The desired outcome is achieved by e.g. adversarial training (one of the components is an adversarial network which forces the encoder to build a representation of the input such that the selected attribute is not deducible from its output) (Goodfellow et al., 2020; Zhang et al., 2018; Lample et al., 2017). Similarly, one can use variational autoencoders, encode the input in the latent probabilistic space and then shift it towards a desired representation (Liu et al., 2020a).

2.4 Gender Bias in Machine Translation

Keeping correct and consistent gender becomes a challenging task not only when translating to a language with rich morphology, but also in natural language processing in general (Sun et al., 2019; Hovy and Spruit, 2016). Stanovsky et al. (2019) proposed an evaluation method for spotting biases in the text and showed that current state-of-the-art models are extensively prone to gender biases. A way to mitigate the effects of such bias is to train a machine translation model which uses additional input in the form of word-level annotations containing information about the subject's gender (Stafanovičs et al., 2020). Another approach is to use a method based on transfer learning (Saunders and Byrne, 2020), utilizing a small set of trusted, gender-balanced examples. This approach gives a strong improvement in gender debiasing with much less computational cost than training from scratch.

Another option is to use methods which modify data directly and remove gender biases from the

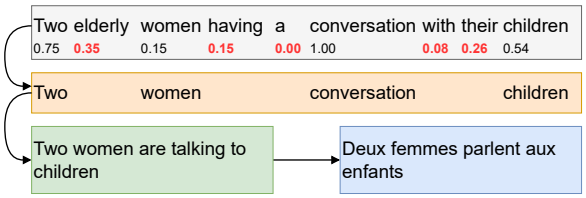

Figure 1: An illustration of a shorter, cascade-generated translation of a sentence with words labeled by significance scores underneath. Steps: 1. Generating scores; 2. Removing words under a threshold; 3. Recovering grammar; 4. Translating. The phrase taken from the NLI dataset (Bowman et al., 2015).

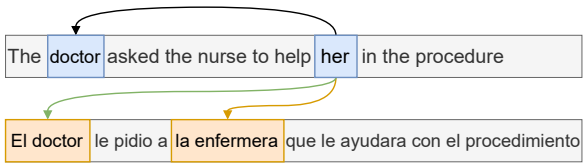

Figure 2: An illustration of gender bias in machine translation. *'doctor'* is mistakenly translated as male profession *'el doctor'* even though the context *'her'* identifies her as feminine. It is caused by wrong coreference alignment between *'her'* and *'nurse'*. The example inspired by Stanovsky et al. (2019).

training data (Zmigrod et al., 2019; Zhao et al., 2018).

3 Method

3.1 Tasks

Let us have the following phrase *'Two elderly women having a conversation with their children'*. Intuitively, we can say that some of the words in the phrase contribute to the overall meaning less than others. For instance, we would say that *'a'* carries considerably less information than *'children'*, or that *'elderly'* is slightly less important than *'women'*. Supposing that we have a scoring function $\mathcal{S} \rightarrow [0, 1]$ which assigns an importance score to each input word in the sentence, we easily identify words that contribute to the meaning the least. It results in a scalable shortening translation if the translation model is capable of accepting and working with such scores as additional inputs. Figure 1 shows an example of shortening machine translation, the first task with constraints we focus on.

Similarly, having the sentence *'The doctor asked the nurse to help her in the procedure'* we can see that the gender of *'doctor'* is not expressed in English.[2] Consequently, when translating to a language in which the surface form of this word differs across genders, the gender has to be inferred from the context. Here, the information about the gender is captured in *'her'*. The goal of the scoring function $\mathcal{S}$ in this context identifies words that aid to disambiguate the translation. Figure 2 presents the problem of gender bias in machine translation, the second task with constraints in the work.

We suppose that solving a task with given constraints can be decomposed to two parts. Formally, we denote these parts as *a constraining task* and *a primary task*. Theoretically, the model trained for solving the constraining task learns features that represent the constraints and is not biased towards formulations exemplified in training data of the primary task (the scoring function $\mathcal{S}$). On the contrary, the model trained for the primary task is not restricted in any way and is able to learn general features of the original task (machine translation). We show the details about constraining tasks for both problems in Section 4.

3.2 Scoring

To obtain the scoring function $\mathcal{S}$, we first train the model for the constraining task. Then, we use an interpreter relying on attribution methods that aims to identify the words that are the most important for explaining the decision of the model trained for the constraining task. These attribution scores are extracted from the constraining task's hidden states through a series of masking procedures.

In particular, we employ the attribution method proposed by De Cao et al. (2020) that seeks to mask the largest possible number of words in the input, while at the same time preserving the output decision obtained from the full input. This means that the interpreter minimizes a loss function comprising two terms: an L_0 term, on the one-hand, which forces the interpreter to maximize the number of masked elements; a divergence term D_*, on the other hand, which aims to diminish the difference between the prediction of the constraining task model given (a) the original input or (b) the masked input. We selected the work of De Cao et al. (2020) over all other approaches, because its main contribution is to approximate the masking process in a fully differentiable loss and thus improve the learning process compared to the other methods.

[2]Note that we focus only on resolving gender bias in sentences where the gender is deducible from the context or from the user input.

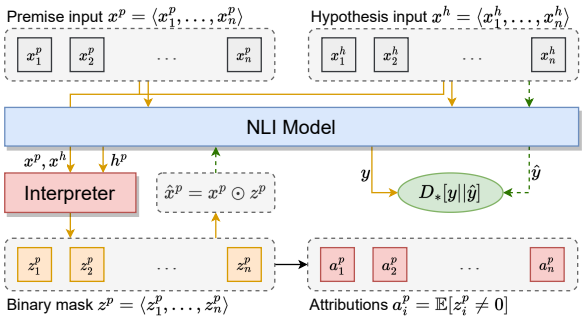

Figure 3: The first pass (yellow plain arrows): A premise and hypothesis are passed to the NLI model. The interpreter takes both text inputs x^p, x^h, and hidden states h^p of the NLI model's encoder. It generates a binary mask z^p which is used to mask x^p, resulting in $\hat{x}^p$. The second pass (green dashed arrows): $\hat{x}^p$ is passed to the NLI model together with the original hypothesis. The divergence D_* minimizes the difference between predicted distributions y and $\hat{y}$ of these two passes.

Having obtained attribution scores, we interpret them as importance scores which reflect the phenomenon defined by the constraints. These scores are then used as additional information for training the model of the primary task.

3.3 Modelling

This additional information in the form of word-level scores can be used in different ways. One option can be to use these scores to scale word embeddings because it has been shown they encode meaning (Conneau et al., 2018; Sileo et al., 2019; Adi et al., 2017). For shortening machine translation, a specific possible direction can be to set a threshold and remove tokens with scores below its value. The subsequent MT model only needs to be pretrained on noisy input (Liu et al., 2020b), or to be created as a cascade of a monolingual denoiser (Lewis et al., 2020) and a vanilla machine translation. Another approach is to integrate scores into textual input of the model (Jain and Berg-Kirkpatrick, 2021).

4 Tasks

We describe two tasks — shortening machine translation and gender bias in machine translation. For the sake of clarity, we denote a constraining task as CT, a primary task as PT and explicitly state the global/local type. Formally, given an autoregressive decoding function F, the constraint C is global when we can formulate the next token prediction x_{n+1} as

$$x_{n+1} = F(x_0, x_1, ..., x_n, C),$$

otherwise the constraint C_i is a local constraint modeled as

$$(x_{n+1}, C_{n+1}) = F(x_0, x_1, ..., x_n, C_n).$$

In other words, global constraints are invariant to the decoding whereas local constraints may change over time.

4.1 Shortening in NMT

PT	Machine Translation
CT	Natural Language Inference
TYPE	Global constraint

Application Generating translation which is shorter than the output from a regular machine translation model is often beneficial in online subtitling (a display system cannot handle the load of incoming subtitles), dubbing (the translation length does not match actor's mouth movements) or in non-isometric machine translation (the pace of the source language is higher than in the target language). This is yearly a part of IWSLT shared tasks (Agarwal et al., 2023).

Description In our setting, we suppose that a model trained for a semantic task is able to learn semantic properties of sentences. Additionally, we assume that word significance within the sentence (defined as word importance) can be estimated by the amount of contribution to the overall meaning of the sentence. This means that removing low-scored words should only slightly change the sentence meaning. We thus use a semantic task (e.g. Natural Language Inference or Paraphrase Identification) as the constraining task. We use the available datasets (Bowman et al., 2015; Williams et al., 2018; Rajpurkar et al., 2016) and train the constraining model. Then, by training an interpreter (De Cao et al., 2020) we obtain a score for each input token. We illustrate the process in Figure 3.

We use these scores as additional information for training the model for the primary task (machine translation), as discussed in Section 3.3.

Evaluation Concerning direct compression across languages, Ive and Yvon (2016) provide a small set of ∼1k parallel sentences extracted from the testset of the WMT 2014 News translation shared task that has been compressed

in the language pair English↔French in both ways. Alternatively, there are several options for monolingual text compression: Text compression data (in domains Europarl, TED, EU bookshop, News) containing English, French, and German (Mallinson et al., 2018); A corpus which contains manual compressions for single and multiple sentences in English (Toutanova et al., 2016); And several other (Cohn and Lapata, 2013; Filippova and Altun, 2013; Clarke and Lapata, 2008).

We evaluate our systems using these datasets and the metrics such as SARI (Xu et al., 2016), Rouge-N or Rouge-L (Lin, 2004). These metrics are used for evaluating sentence simplification or summarization systems. For the translation quality we use common measures such as BLEU (Papineni et al., 2002), model-based COMET (Rei et al., 2020) or character n-gram F-score chrF (Popović, 2015). For baselines, we employ pretrained models for sentence summarization (Zhang et al., 2019; Rothe et al., 2020).

4.2 Gender Bias in NMT

PT	Machine Translation
CT	Coreference
TYPE	Global constraint

Application Consistent translation of words that exist in different genders, and reducing the impact of gender bias is useful in situations when the target language has comparably richer morphology than the source language. A possible usage is also in the context of translating dialogues: Given a limited history of a dialogue, the translation system has to correctly assess the gender of speakers. Stanovsky et al. (2019) present a challenge set and evaluation protocol for the analysis of gender bias in machine translation, which was used as a test suite of WMT20 by Kocmi et al. (2020).

Description In our setting, we explore several options for the constraining task but we suppose that training a model for detecting coreference allows to detect words which are important for sustaining the correct gender within or across sentences. We use the available datasets (Bamman et al., 2020; Webster et al., 2018) to train the constraining model – given a sentence (or multiple sentences) and a noun that is contained in the input, the model predicts the gender of the noun, e.g. *'The doctor asked the nurse to help her. || doctor → feminine'*. Thanks to the coreference resolution datasets we can correctly derive gender for gold labels. Similarly to shortening machine translation, we train the interpreter (De Cao et al., 2020) for generating importance scores of the input tokens which contribute the most to the decision of choosing the word gender of the given noun. We use these tokens in the next step as additional information for the model of the primary task – we pretrain an MT model with the form of the input *'sentence || noun || word that helps to identify the gender of the noun → sentence translation'*. For the inference, we ask the constraining model for the word that helps with the gender choice and use it for the subsequent MT model such as *'The doctor asked the nurse to help her. || doctor || her → La doctora...'*, supposing that *'her'* is the token which achieves the highest importance score. The model should select the correct translation *'la doctora'* which is a feminine noun.

Evaluation Zhao et al. (2018) publish a dataset called WinoBias that contains ∼3k sentences. Annotators were asked to describe situations where entities interact in plausible ways. The templates were selected to be challenging and designed to cover cases requiring semantics and syntax separately. Similarly, Stanovsky et al. (2019) present WinoMT, a dataset containing ∼4k instances, equally balanced between male and female genders, as well as between stereotypical and non-stereotypical gender-role assignments.

We use regular machine translation systems for baselines and evaluate the quality of our systems using simple metrics such as accuracy or F_1-measure.

5 Experiments

This section describes preliminary results that are related to the first task: Shortening in NMT. As we introduced in the description part of Section 4.1, we first aim to detect important words in sentences. We conduct experiments and show the results in Section 5.1. Second, having the importance scores, our goal is to use these scores to make MT models aware of the importance of the tokens on their input. For this, we examine scaling word embeddings because we suppose that they encode meaning. We show results of such experiments in Section 5.2.

Note that all experiments are related to the first task and we have not conducted any experiments for the second task, gender bias in NMT, yet.

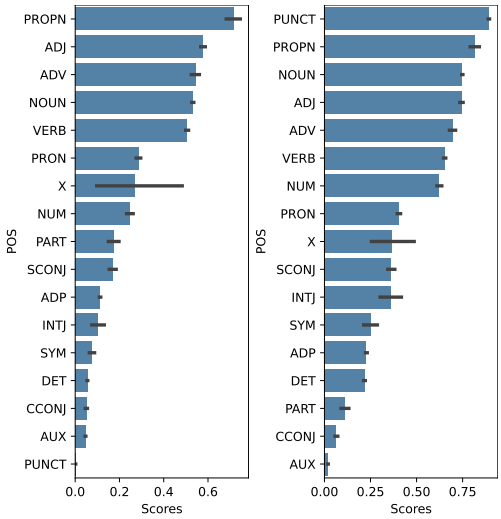

Figure 4: Average scores for each POS category for the NLI model (left) and PI model (right).

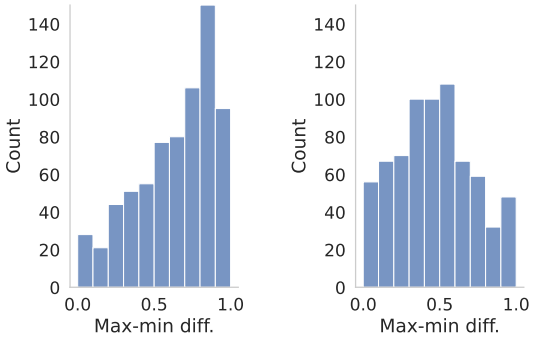

Figure 5: The NLI model (left), PI model (right) and the distribution of differences between the maximal and minimal value for each token.

	NLI Model		PI Model		
Depth	**Avg**	**Std**	**Avg**	**Std**	**Count**
1	**0.52**	0.35	**0.64**	0.31	9424
2	**0.36**	0.36	**0.53**	0.39	27330
3	**0.23**	0.31	**0.40**	0.35	26331
4	0.22	0.31	0.33	0.36	7183
5	0.22	0.30	0.35	0.35	1816

Table 1: Importance scores of tokens for each depth in syntactic trees. Stat. significant differences between the current and next row are bolded ($p < 0.01$).

5.1 Word importance

The following text is a selection of results presented in Javorský et al. (2023).[3] We follow the methodology from Section 3 and construct a pipeline as shown in Figure 3. An example of our importance scores is displayed in Figure 1. In this section, we show properties that the importance scores have, evaluated on the SNLI validation set (Bowman et al., 2015). Recall that we independently train two semantic models: One for Natural Language Inference (NLI) and one for Paraphrase Identification (PI).

5.1.1 Content Words are More Important

We first examine the scores that are assigned to content and functional words. We compute the average score for each POS tag (Zeman et al., 2022) and display the results in Figure 4. For both models, Proper Nouns, Nouns, Pronouns, Verbs, Adjectives and Adverbs have leading scores. Determiners, Particles, Symbols, Conjunctions, Adpositions are scored lower. We observe an inconsistency of the PI model scores for Punctuation. We suppose this reflects idiosyncrasies of the PI dataset: Some items contain two sentences within one segment, and these form a paraphrase pair only when the other segment also consists of two sentences. Therefore, the PI model is more sensitive to Punctuation than expected. We also notice the estimated importance of the X category varies widely, which is expected since this category is, based on its definition, a mixture of diverse word types. Overall, these results fulfil our requirement that content words achieve higher scores than function words.

5.1.2 Word Significance is Context-Dependent

We then question the ability of the interpreter to generate context-dependent attributions, contrasting with purely lexical measures such as TF-IDF. To answer this question, we compute the distribution of differences between the lowest and highest scores for words having at least 100 occurrences in the training and 10 in the validation data, excluding tokens containing special characters or numerals. The full distribution is plotted in Figure 5.

Scores extracted from both models report increased distribution density towards larger differences, confirming that significance scores are not lexicalized, but instead strongly vary according to the context for the majority of words. The greatest difference in scores for PI model is around 0.5, the analysis of the NLI model brings this difference even more towards 1. We explain it by the nature of the datasets: It is more likely that the NLI model's decision relies mostly on one or on a small group of words, especially in the case of contradictions.

[3] Note that several parts of this section are cited as they appear in the original paper.

5.1.3 Important Words are High in the Tree

Linguistic theories differ in ways of defining dependency relations between words. One established approach is motivated by the 'reducibility' of sentences (Lopatková et al., 2005), i.e. gradual removal of words while preserving the grammatical correctness of the sentence. In this section, we study how such relationships are also observable in attributions. We collected syntactic trees of input sentences with UDPipe (Straka, 2018),[4] which reflect syntactic properties of the Universal Dependencies format (Zeman et al., 2022).[5] When processing the trees, we discard punctuation and compute the average score of all tokens for every depth level in the syntactic tree. We display the first 5 depth levels in Table 1.

We can see tokens closer to the root in the syntactic tree obtain higher scores on average. We measure the correlation between scores and tree levels, resulting in -0.31 Spearman coefficient for the NLI model and -0.24 for the PI model. Negative coefficients correctly reflect the tendency of the scores to decrease in lower tree levels. It thus appears that attributions are well correlated with word positions in syntactic trees, revealing a relationship between semantic importance and syntactic depth.

5.2 Downscaling Word Embeddings

We assume that down-scaling word embeddings makes the input signal weaker. For generating new embeddings, we multiply embeddings of tokens that we aim to alter by a scaling factor α. We denote two operations: (a) `SCALE` if $\alpha \in [0, 1]$ or (b) `BINARY` if $\alpha = 0$.

5.2.1 Experiment Setup

We selected the family of `Helsinki-NLP` models to conveniently analyze the behavior of translation models across different language pairs (Tiedemann and Thottingal, 2020). We study en→{fr,cs,de,zh} to cover various language families. Note that Transformer models contain two layers of embeddings: *word* and *positional*. We apply our strategies to both of them and to all subwords that belong to one word. We use a subset of the `MS COCO` dataset (Lin et al., 2014), with the size of 677 sentences. We evaluate in two modes:

[4] https://lindat.mff.cuni.cz/services/udpipe/

[5] UD favors relations between content words, function words are systematically leaves in the tree. However, having function words as leaves better matches our perspective of information importance flow, unlike in Gerdes et al. (2018).

	Spearman rank correlation			
Direction	**en→fr**	**en→cs**	**en→de**	**en→zh**
en→fr	*1.00*	*0.71*	*0.90*	0.62
en→cs	*0.71*	*1.00*	*0.90*	*0.88*
en→de	*0.90*	*0.90*	*1.00*	*0.83*
en→zh	0.62	*0.88*	*0.83*	*1.00*

Table 2: Spearman rank correlation coefficients: *Italic* and underlined values are statistically significant with $p < 0.05$ and $p < 0.01$, respectively.

Encoder To know what the impact of embedding modifications is at the encoder output level, we take the `MAX` pooling of the encoder output and compute the cosine similarity between the output from the encoder before and after making modifications to the model input.

Decoder We also study the impact of the changes in the input on translation. Following Fadaee and Monz (2020), we use the Levenshtein distance (Levenshtein et al., 1966) for evaluating the translation of the distorted input compared to the original translation. The computation of the Levenshtein distance is based on a set of editing operations comprising insertion, deletion, and substitution.

5.2.2 Encoder Results

`SCALE` The effects of embedding down-scaling on the encoder output are displayed in Figure 6. We can see that the curves differ across language pairs: French, Czech, and German seem to be more sensitive to this type of modification, Chinese, on the other hand, is minimally influenced up to the scaling factor of 0.4. In other words, 'graying out' an English source word affects the encoder in en→zh faster than other languages: the cosine similarity starts to decrease at around α of 0.7 and at already 0.5 reaches its lowest level (0.985). German also exposes an anomaly: the curve is not monotone. Overall, we conclude that the information stored in contextualized embeddings is differently distributed among languages, although in absolute terms, the max-pooled contextualized embeddings are not affected much when one word in the sentence is fully 'grayed out'.

`BINARY` In Figure 7, we can see in more detail the impact of zeroing some embeddings from the encoder on the encoder output. We observe that the cosine similarity in Chinese is generally lower than in other languages. The results also show that nouns are the most important in the encoder representation and numbers are the least, which is consistent for all target languages.

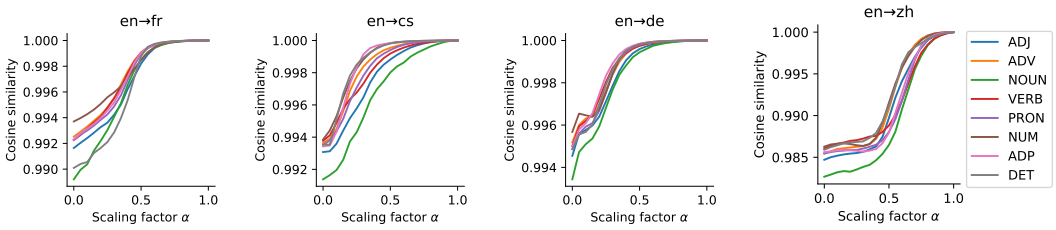

Figure 6: The cosine similarity of the encoder output after MAX pooling given the original and modified input. The scaling factor is sampled with step 0.05.

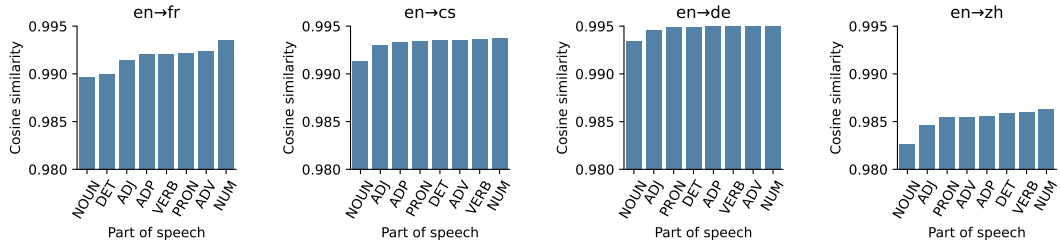

Figure 7: The cosine similarity of the encoder output after MAX pooling given the original and fully-zeroed input.

	Cosine similarity			
Pooling	**en→fr**	**en→cs**	**en→de**	**en→zh**
MAX	0.95±0.01	0.95±0.01	0.97±0.01	0.94±0.01
MEAN	0.87±0.22	0.95±0.05	0.89±0.17	0.92±0.12

Table 3: The cosine similarity of two random sentences when performing the MAX or MEAN pooling.

Even though the absolute values for POS categories are different across languages, we notice that the relative ordering is remarkably similar. Therefore, we compute the Spearman rank correlation for each pair of models and display the results in Table 2. We observe strong correlations between all models except the pair of models en→fr and en→zh. These findings confirm that encoders learn features of the source language more or less independently of the target language.

Figure 6 and Figure 7 suggest that the word embedding space is very narrow: The change in one word does not influence the sentence representation by a lot. To confirm this, we examine the cosine similarity of two random sentences. We perform the MAX and MEAN pooling on the encoder output.

Table 3 presents the results. We observe three properties. First, the MAX pooling makes the space extremely narrow: The similarity of two random sentences ranges from 0.94 (Chinese) to 0.97 (German), with a low variance everywhere. This observation is remarkably consistent across all four target languages. Second, the MEAN pooling reaches different values in different languages, ranging from 0.87 for French to 0.95 for Czech. Third, the difference between MAX and MEAN varies across languages: French exhibits a bigger difference than Czech. This observation is interesting since Czech is a language with richer morphology.

5.3 Decoder Results

SCALE Fadaee and Monz (2020) show that making minor changes to the input sentence causes major differences in the translation. We test how many times the model changes the prediction during down-scaling of word embeddings. For this experiment, we gradually apply the scaling factor $\alpha \in \{0, 0.01, ..., 0.99, 1.0\}$ and count the number of different translations given two adjacent scaling factors. In other words, we count how many times the output changes with the scaling factor α on the path from zero to one with step 0.01.

Table 4 displays the average and standard deviation of these counts. Changes in embeddings of nouns and verbs influence the translation the most, which is intuitive. Pronouns, numbers, and adverbs have the slightest impact. Interestingly, however, altering embeddings of determiners (articles) is almost as influential as content-rich words such as adjectives. Additionally, there are more than 20 changes when modifying nouns out of 100 different translations for all languages. This indicates that down-scaling embeddings to zero is not smooth: It critically affects the translation. This non-smoothness is striking in en→zh while gray-

	Number of changes							
Direction	**NOUN**	**VERB**	**ADJ**	**ADV**	**PRON**	**NUM**	**ADP**	**DET**
en→fr	22.48±13.12	13.38±12.82	8.35±10.00	5.36±7.32	4.51±5.24	5.01±5.21	11.41±10.76	8.24±8.26
en→cs	27.65±17.57	15.77±15.78	10.12±12.54	6.10±7.39	6.30±7.16	5.25±6.11	11.71±12.72	8.21±9.56
en→de	21.96±14.68	13.59±13.35	8.17±9.75	5.79±6.67	5.65±6.61	5.18±4.79	12.16±11.54	7.87±8.27
en→zh	41.95±20.43	24.70±20.33	16.34±17.95	10.00±11.45	9.09±10.08	9.90±9.67	22.10±19.69	15.24±15.82

Table 4: Each cell contains the mean ± standard deviation of the number of translation pairs for which the translation differs, taking the pairs with scaling factors α_1, α_2 where $|\alpha_1 - \alpha_2| = 0.01$.

	Levenshtein distance on words							
Direction	**NOUN**	**VERB**	**ADJ**	**ADV**	**PRON**	**NUM**	**ADP**	**DET**
en→fr	4.77±3.20	3.50±2.55	3.27±2.30	3.34±2.47	3.34±2.51	3.19±2.40	3.52±2.80	3.30±4.05
en→cs	3.92±2.68	3.81±13.31	2.83±2.26	3.35±2.76	3.62±2.84	3.39±2.44	3.65±3.54	2.22±2.74
en→de	4.69±15.39	3.55±2.62	3.10±2.49	3.18±2.72	3.92±2.69	3.12±2.77	3.72±3.14	3.69±2.98

	Levenshtein distance on characters							
Direction	**NOUN**	**VERB**	**ADJ**	**ADV**	**PRON**	**NUM**	**ADP**	**DET**
en→fr	21.26±14.91	15.48±10.91	15.64±9.96	14.04±10.61	12.86±10.82	11.65±9.80	14.13±12.06	12.85±18.61
en→cs	15.79±10.75	14.74±28.06	13.31±8.85	13.72±11.84	13.74±11.47	12.46±9.98	12.80±16.56	9.13±11.03
en→de	20.64±65.36	15.03±11.09	14.64±10.98	13.77±11.70	15.32±11.34	12.49±11.23	13.83±12.32	13.68±12.50
en→zh	15.00±43.46	9.27±8.49	10.91±23.90	10.87±26.36	9.37±19.88	10.26±18.64	10.07±22.70	8.52±14.07

Table 5: The word (upper table) and character (lower table) Levenshtein distance of the original translation and the translation when zeroing embeddings. Each cell presents mean ± standard deviation.

ing out a noun in the sentence: in 100 steps, there are on average 40 changes of the target sentence.

BINARY We compute the word and character Levenshtein distances and display statistics in Table 5. The results suggest that masking embeddings almost equally affects the number of changed words in the translation, even for the least meaning-bearing words such as determiners. Specifically, zeroing out one word in the source leads to an edit distance of 3–4 on average. We observe the translation fluctuation is less noticeable for Czech for both measures. Furthermore, we notice a very high variance in the word distance for nouns in German and verbs in Czech. While we assumed that this could have been caused by reordering in the relatively free word order languages, an inspection of the data revealed that the model began to generate repeated tokens.

6 Conclusion

In this thesis proposal, we provide an overview of the current state of the research in this field: Transfer learning methods, model interpretability techniques and two selected problems, shortening machine translation and gender bias in machine translation, and their description. We present an innovative way how to train models for tasks with constraints. We anticipate training two models: a *constraining model* that learns the constraint features of the task, and a *general model* that learns general features of the task. We suppose that analyzing the constraining model by an attribution method we obtain scores that correlate well with constraint features learn by this model. We then use these score and apply them to the general model.

In our experiments, we show that importance scores that we acquire have desired and meaningful properties: Content words are more important, scores are context-dependent and words closer to the root in its sentence syntactic tree are more important on average. We then show the analysis of downscaling word embeddings of Transformer-based translation models and we observe models' behavior on two levels: the encoder and decoder output. For the encoder, we show that continuous scaling of embeddings affects different language pairs slightly differently but the relations between POS categories are similar, showing strong Spearman rank correlation. For the decoder, we observe that gradual transition of embeddings affects the translation in a not very smooth way (we see many changes for every language), and that graying out even the least meaning-bearing words can have a significant impact on the translation.

Future Work First, we anticipate finishing the task of shortening in NMT. We explore different ways to use the scores as described in Section 3.3: (a) We either implement a pipeline that removes low-scored tokens and the output is recovered using models trained on noisy input; Or (b) we use the scores directly as additional model's input. We describe the overall approach in Section 4.1.

Second, we aim at using the proposed method to address gender bias in machine translation as

presented in Section 4.2. We train the constraining model for coreference classification and analyze it with an interpreter. We use the word that achieves the highest score for the subsequent MT model which is pretrained in a way that accepts such information. We believe that this approach can help mitigate gender bias in machine translation since we elevate the importance of certain words on the input which are usually ignored because such situations are not frequent enough for the MT model to learn.

Note that the choice of the constraining and primary task is not always trivial and requires the knowledge about the available datasets and the relations between them in terms of constraints.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2017. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. In *International Conference on Learning Representations*. International Conference on Learning Representations, ICLR.

Milind Agarwal, Sweta Agrawal, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, Mingda Chen, William Chen, Khalid Choukri, Alexandra Chronopoulou, Anna Currey, Thierry Declerck, Qianqian Dong, Kevin Duh, Yannick Estève, Marcello Federico, Souhir Gahbiche, Barry Haddow, Benjamin Hsu, Phu Mon Htut, Hirofumi Inaguma, Dávid Javorský, John Judge, Yasumasa Kano, Tom Ko, Rishu Kumar, Pengwei Li, Xutai Ma, Prashant Mathur, Evgeny Matusov, Paul McNamee, John P. McCrae, Kenton Murray, Maria Nadejde, Satoshi Nakamura, Matteo Negri, Ha Nguyen, Jan Niehues, Xing Niu, Atul Kr. Ojha, John E. Ortega, Proyag Pal, Juan Pino, Lonneke van der Plas, Peter Polák, Elijah Rippeth, Elizabeth Salesky, Jiatong Shi, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Yun Tang, Brian Thompson, Kevin Tran, Marco Turchi, Alex Waibel, Mingxuan Wang, Shinji Watanabe, and Rodolfo Zevallos. 2023. FINDINGS OF THE IWSLT 2023 EVALUATION CAMPAIGN. In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 1–61, Toronto, Canada (in-person and online). Association for Computational Linguistics.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France. European Language Resources Association.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Stevo Bozinovski and Ante Fulgosi. 1976. The influence of pattern similarity and transfer learning upon training of a base perceptron b2. In *Proceedings of Symposium Informatica*, volume 3, pages 121–126.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Trevor Cohn and Mirella Lapata. 2013. An abstractive approach to sentence compression. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):1–35.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. How do decisions emerge across layers in neural models? interpretation with differentiable masking. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Marzieh Fadaee and Christof Monz. 2020. The unreasonable volatility of neural machine translation models. In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 88–96, Online. Association for Computational Linguistics.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, L Alfonso Ureña-López, and José Ignacio Abreu Salas. 2021. A survey on bias in deep nlp. *Applied Sciences (2076-3417)*, 11(7).

Kim Gerdes, Bruno Guillaume, Sylvain Kahane, and Guy Perrier. 2018. SUD or surface-syntactic Universal Dependencies: An annotation scheme near-isomorphic to UD. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 66–74, Brussels, Belgium. Association for Computational Linguistics.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144.

Jiatao Gu, Yong Wang, Yun Chen, Victor O. K. Li, and Kyunghyun Cho. 2018. Meta-learning for low-resource neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3622–3631, Brussels, Belgium. Association for Computational Linguistics.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. 2019. Towards a deep and unified understanding of deep neural models in nlp. In *International conference on machine learning*, pages 2454–2463. PMLR.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Julia Ive and François Yvon. 2016. Parallel sentence compression. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, page 1503–1513, Osaka, Japan.

Aashi Jain and Taylor Berg-Kirkpatrick. 2021. An empirical study of extrapolation in text generation with scalar control. *arXiv preprint arXiv:2104.07910*.

Dávid Javorský, Ondřej Bojar, and François Yvon. 2023. Assessing word importance using models trained for semantic tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8846–8856, Toronto, Canada. Association for Computational Linguistics.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Tom Kocmi, Tomasz Limisiewicz, and Gabriel Stanovsky. 2020. Gender coreference and bias evaluation at WMT 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 357–364, Online. Association for Computational Linguistics.

Surafel Melaku Lakew, Mattia Di Gangi, and Marcello Federico. 2019. Controlling the output length of neural machine translation. In *Proceedings of the 16th International Conference on Spoken Language Translation*, Hong Kong. Association for Computational Linguistics.

Guillaume Lample, Neil Zeghidour, Nicolas Usunier, Antoine Bordes, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Fader networks: manipulating images by sliding attributes. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 5969–5978.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020a. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Markéta Lopatková, Martin Plátek, and Vladislav Kuboň. 2005. Modeling syntax of free word-order languages: Dependency analysis by reduction. In *International Conference on Text, Speech and Dialogue*, pages 140–147. Springer.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6470–6479.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Dominik Macháček, Matúš Žilinec, and Ondřej Bojar. 2021. Lost in interpreting: Speech translation from source or interpreter? In *Proceedings of INTERSPEECH 2021*, pages 2376–2380, Baxas, France. ISCA.

Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55(8):1–42.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2018. Sentence compression for arbitrary languages via multilingual pivoting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2453–2464, Brussels, Belgium. Association for Computational Linguistics.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: contextualized word vectors. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6297–6308.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):4381.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Danielle Saunders and Bill Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7724–7736, Online. Association for Computational Linguistics.

Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the flow: Information bottlenecks for attribution. In *International Conference on Learning Representations*.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Composition of sentence embeddings: Lessons from statistical relational learning. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 33–43, Minneapolis, Minnesota. Association for Computational Linguistics.

Artūrs Stafanovičs, Toms Bergmanis, and Mārcis Pinnis. 2020. Mitigating gender bias in machine translation with target gender annotations. In *Proceedings of the Fifth Conference on Machine Translation*, pages 629–638, Online. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Sandeep Subramanian, Adam Trischler, Yoshua Bengio, and Christopher J Pal. 2018. Learning general purpose distributed sentence representations via large scale multi-task learning. In *International Conference on Learning Representations*.

Umut Sulubacak, Ozan Caglayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34:97–147.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Sho Takase and Naoaki Okazaki. 2019. Positional encoding to control output sequence length. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3999–4004, Minneapolis, Minnesota. Association for Computational Linguistics.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – building open translation services for the world. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

Kristina Toutanova, Chris Brockett, Ke M. Tran, and Saleema Amershi. 2016. A dataset and evaluation metrics for abstractive compression of sentences and short paragraphs. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 340–350, Austin, Texas. Association for Computational Linguistics.

Gido M van de Ven and Andreas S Tolias. 2019. Three scenarios for continual learning. *arXiv preprint arXiv:1904.07734*.

Shikhar Vashishth, Shyam Upadhyay, Gaurav Singh Tomar, and Manaal Faruqui. 2019. Attention interpretability across nlp tasks. *arXiv preprint arXiv:1909.11218*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kellie Webster, Marta Recasens, Vera Axelrod, and Jason Baldridge. 2018. Mind the GAP: A balanced corpus of gendered ambiguous pronouns. *Transactions of the Association for Computational Linguistics*, 6:605–617.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv e-prints*, pages arXiv–2206.

Marion Weller-di Marco and Alexander Fraser. 2022. Findings of the WMT 2022 shared tasks in unsupervised MT and very low resource supervised MT. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 801–805, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Jie Yang, Yue Zhang, and Fei Dong. 2017. Neural word segmentation with rich pretraining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 839–849, Vancouver, Canada. Association for Computational Linguistics.

Daniel Zeman et al. 2022. Universal dependencies 2.10. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization.

Ruiqing Zhang, Chuanqiang Zhang, Zhongjun He, Hua Wu, and Haifeng Wang. 2020. Learning adaptive segmentation policy for simultaneous translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2280–2289, Online. Association for Computational Linguistics.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2023. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 943–956, Dubrovnik, Croatia. Association for Computational Linguistics.

Zhirui Zhang, Shujie Liu, Mu Li, Ming Zhou, and Enhong Chen. 2018. Bidirectional generative adversarial networks for neural machine translation. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 190–199, Brussels, Belgium. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.